StRDAN: Synthetic-to-Real Domain Adaptation Network for Vehicle Re-Identification

Sangrok Lee[1], Eunsoo Park[1], Hongsuk Yi[2], and Sang Hun Lee*[3]

[1]MODULABS
srl@modulabs.ai, es.park@modulabs.co.kr
[2]Korea Institute of Science and Technology Information
hsyi@kisti.re.kr
[3]Kookmin University
shlee@kookmin.ac.kr

Abstract

Vehicle re-identification aims to obtain the same vehicles from vehicle images. It is challenging but essential for analyzing and predicting traffic flow in the city. Although deep learning methods have achieved enormous progress in this task, requiring a large amount of data is a critical shortcoming. To tackle this problem, we propose a novel framework called Synthetic-to-Real Domain Adaptation Network (StRDAN), which is trained with inexpensive large-scale synthetic data as well as real data to improve performance. The training method for StRDAN is combined with domain adaptation and semi-supervised learning methods and their associated losses. StRDAN shows a significant improvement over the baseline model, which is trained using only real data, in two main datasets: VeRi and CityFlow-ReID. Evaluating with the mean average precision (mAP) metric, our model outperforms the reference model by 12.87% in CityFlow-ReID and 3.1% in VeRi.

1. Introduction

Vehicle re-identification (Re-ID) aims to identify the same vehicles that are captured by various cameras. It is an essential technology for analyzing and predicting traffic flow in a smart city and uses visual appearance-based Re-ID methods in general. However, Vehicle Re-ID is challenging for two reasons. First, different lighting and complex environments create difficulties with appearance-based vehicle Re-ID. Also, if the vehicle is captured using different cameras, large variations in appearance will be produced. Secondly, different vehicles can be very similar to each other visually when they are in the same category.

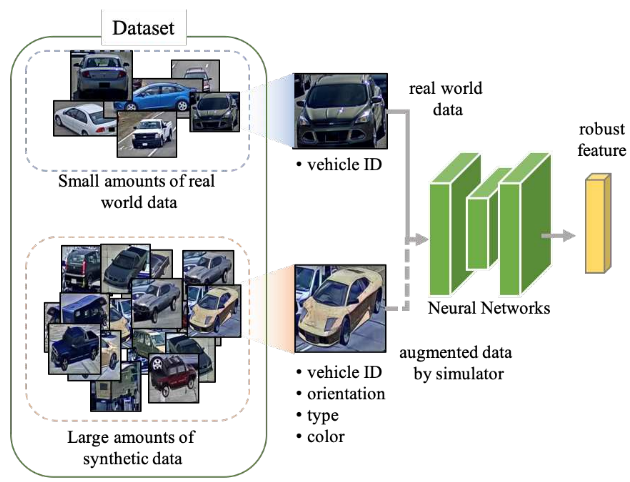

Figure 1. Synthetic-to-Real Domain Adaptation method to improve the performance of vehicle Re-ID. In general, it is difficult to get labels for real data, but it is easy for synthetic data.

Deep learning methods [24, 9, 17] are used to tackle this complex vehicle Re-ID task and achieve significant progress. They extract features using deep learning networks and distinguish vehicles by comparing the distances between their features. However, requiring a large amount of data to improve performance is a drawback of deep learning. The reported result [33] shows that the more training data a model has, the better performance it makes. Data from the wild environment need a heavy workload of annotation. Many studies have attempted to use inexpensive synthetic data to replace real data. Such research is called

*corresponding author

domain adaptation.

In this paper, we explore how to improve model performance using inexpensive synthetic data (see Fig.1). First, we have adopted an adversarial domain adaptation approach [3] in which a neural network learns features that are as discriminative as possible for the main classification task on the real data, while, at the same time, learning indistinguishable features between real and synthetic data [1] [4]. To implement this idea, we introduce a domain discrimination layer and associated cross-entropy loss to train the network indiscriminative for both domains. Secondly, to exploit the specific labels in synthetic data such as color, type, and orientation, we have also adopted semi-supervised learning methods. Since these labels exist only in synthetic data, a semi-supervised learning approach that can handle unlabeled data is applicable to improve the performance. In training, classification losses for the exclusive labels are selectively applied depending on the data domain [31]. Our model trained the real and synthetic data of the AI City Challenge using the domain-adaptation and semi-supervised learning approach was 12.87% better than the baseline model that was trained with only real data.

In this work, we propose a novel framework named StRDAN, standing for Synthetic-to-Real Domain Adaptation Network. Our major contribution is threefold:

- StRDAN is trained with inexpensive large-scale synthetic data as well as real data to improve the performance.
- A new training approach for StRDAN is proposed, which is combined with domain adaptation and semi-supervised learning methods and corresponding losses.
- StRDAN shows a significant improvement over the baseline model in two main data sets: VeRi [14] and CityFlow-ReID [26].

2. Related Work

In this section, we review the prior works from two aspects: vehicle Re-ID and the domain adaptation method with synthetic data.

Vehicle Re-ID: Vehicle Re-ID methods generally have two characteristics: contrastive loss and spatio-temporal feature. First, in terms of contrastive loss, prior works [13, 14, 15] proposed methods that use constrative loss in forms of siamese network, triplet loss, and metric learning. Liu *et al*. [14] also introduced the VeRi dataset for the first large-scale vehicle Re-ID benchmark. Second, spatio-temporal feature is the key to performance improvement. Vehicle Re-ID task achieved a huge progress using the spatio-temporal features. Tan *et al*. [23] uses spatial-temporal features for multi-camera vehicle tracking and vehicle Re-ID, and their method proved by winning The AI City Challenge 2019[18]. Shen *et al*. [21] proposed a two-stage framework for matching visual appearance and an LSTM based path inference mechanism.

Domain Adaptation with Synthetic Data: To overcome the lack of data, Zhou *et al*. [36, 37] proposed a method that improves the Re-ID performance by augmenting various viewpoint vehicle images with Generative Adversarial Networks (GAN). There is also a method to deal with inconsistency in the distribution of different data sources. When deploying the well-trained model directly to a new dataset, the performance drops significantly due to the differences among datasets named domain bias. Peng *et al*. [19] proposed a domain adaptation framework to address this problem, which contained an image-to-image translation network and an attention-based feature learning network. We can use VehicleX[30] simulator to leverage synthetic data and domain randomization to overcome the reality gap [27, 28]. Liu *et al*. [11] also proposed a domain adaptation method. However, they only considered real-to-real domain adaptation. The recent vehicle Re-ID research [25] proposed PAMTRI uses synthetic data to improve the performance and have a similar architecture with ours. Compared to PAMTRI that requires additional effort to get vehicle pose and label for real data, our StRDAN uses domain adaptation to utilize synthetic data and adopts semi-supervised learning that doesn't need to extra annotation workload. Our method is simple and easy to train.

3. Synthetic-to-Real Domain Adaptation Network (StRDAN)

3.1. Datasets

In this work, we developed a neural network using the real and synthetic vehicle datasets provided for the Track 2 of the 2020 AI City Challenge. The real dataset is the CityFlow-reID dataset, which is a subset of CityFlow made available for the Track 2 challenge and consists of 56,277 images for 666 unique vehicles collected from 40 cameras. Here, 36,935 images from 333 vehicle identities are provided for training, while 18,290 images from the other 333 identities are given for testing. The remaining 1052 images from the same identities in the test set are provided as query data.

The synthetic vehicle dataset consists of 192,150 images from 1,362 distinct vehicles created using a large-scale synthetic dataset generator called VehicleX [30] to form an augmented training set. The synthetic dataset has not only the vehicle ID but also additional information such as the color, type, and orientation of an object, whereas the real dataset has only the vehicle ID. Here, vehicles are distinguished into 12 colors and 11 types. The orientation is represented by a rotation angle on the horizontal plane in the range of [0, 360).

We also trained and evaluated our model using the VeRi real dataset [8] and the City Challenge synthetic data to examine the validity and robustness of our approach. The Veri dataset contains over 50,000 images of 776 vehicles captured by 20 cameras. The training set contains 37,781 images of 576 vehicles, while the testing set contains 11,579 images of 200 vehicles. We don't VeRi has additional labels that are color and type.

3.2. Overall Architecture

The overall architecture of the proposed synthetic-to-real domain adaptation network (StRDAN) is shown in Figure 2. The model consists of a backbone network for feature extraction and multiple fully connected (FC) softmax layers for classification. Input images are sampled in batch in equal numbers from the real and synthetic datasets. For a mini-batch, *n* different vehicle identities are chosen from the real and synthetic datasets, respectively, then *m* samples are randomly selected from the images of the chosen identities. Therefore, each batch contains $2 \times n \times m$ images.

The backbone network extracts a highly-abstracted feature vector (*dim* = 2048) from an input image. Conceptually, any convolution neural network designed for image classification can be used as a backbone network. In prior work, various CNN networks, such as VGG-CNN-M-1024 [2], MobileNet [8], ResNet [6], have been adopted as a backbone for vehicle Re-ID. In the proposed StRDAN, ResNet-50 has been selected as a backbone network. The feature map extracted by the backbone network is flattened and fed into various FC softmax layers for classification of vehicle id, real or synthetic, color, type, and orientation. The outputs are fed into five cross-entropy loss functions and one triplet loss function. Our model is trained in an end-to-end manner by updating the parameters in the network to reduce the total loss, which is a combination of the cross-entropy losses and the triplet loss.

3.3. Key Features

Adversarial Domain Adaptation. An annotated dataset is essential for supervised learning of a deep neural network. However, collecting and manually annotating a large amount of data is a time consuming and expensive task. To overcome this problem, an approach to generate automatically labeled data using a graphic simulator has been introduced. In the AI City Challenge, a synthetic vehicle dataset created using VehicleX is provided to overcome the lack of real data. However, the synthetic data has similar but different distributions, compared to the real data. Therefore, it is necessary to train a neural network to be predictive of the classification task, but uninformative as to the domain of the input.

We adopted the adversarial domain adaptation approach in which a neural network learns features that are as discriminative as possible for the main classification task on the real domain and , at the same time, as indistinguishable as possible between the real and synthetic domains [1] [4]. To implement this idea, we introduced a domain discrimination layer and its associated cross-entropy loss to make the network be trained indiscriminative to the two domains. Also, to train the network more discriminative for vehicle identities and shape signatures, we introduced not only a vehicle-id classification layer and its associated cross-entropy loss but also a triplet loss.

Semi-supervised Learning. The synthetic data has labels such as vehicle type, color, and orientation unlike the real data. We use the labels as multi-task learning to improve generalization performance of all the tasks [32]. In this case, various approaches for semi-supervised learning can be introduced to improve learning accuracy because semi-supervised learning basically combines a small amount of labeled data with a large amount of unlabeled data during training. In Zhai *et al*. 's work [31], they create artificial labels for both unlabeled and labeled data and utilize them in training. Their approach inspired us an idea to use joint and disjoint labels between real and synthetic data for improving the performance. Here, joint labels attached to both real and synthetic data are vehicle ID and domain (real or synthetic), while disjoint labels attached to only synthetic data are vehicle type, color, and orientation. As shown in Figure 2, the losses are also classified into joint and disjoint losses, which are associated with joint and disjoint labels, respectively. The triplet loss is classified as a joint loss because the vehicle id contributes to distinguishing batch images into anchor, positive and negative images.

The semi-supervised learning approach we consider in this paper has a learning objective in the following form:

$$\min_{\theta} \ \mathcal{L}_{joint}(\theta) + w\mathcal{L}_{disjoint}(\theta), \tag{1}$$

where $\mathcal{L}_{joint}$ is the joint loss defined in both real and synthetic domains and $\mathcal{L}_{disjoint}$ is the disjoint loss defined in the synthetic domain. θ is parameters of the network. In the next section, we will describe the losses in more detail.

4. Loss Function

4.1. Joint Losses

Vehicle ID. A cross-entropy loss following the softmax function is the most common loss in image classification. The cross-entropy loss of the vehicle ID classifier, L_{id}, is represented as follows:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}), \tag{2}$$

where N denotes the number of images in a mini-batch, C represents the number of classes, y_{ij} is the j^{th} element

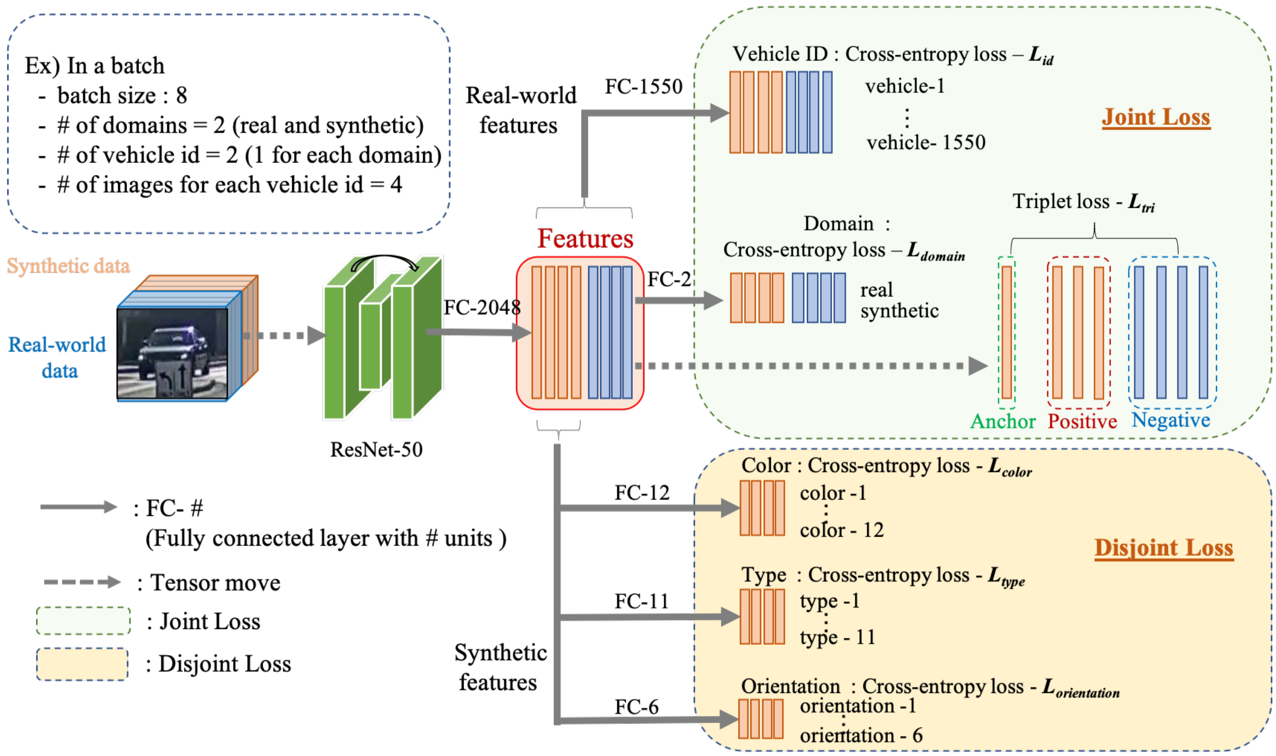

Figure 2. The architecture of the proposed synthetic-to-real domain adaptation network (StoRDAN) that consists of a ResNet-50 backbone for fearure extraction and five fully-connected softmax layers for classification, trained using the joint and disjoint losses between synthetic and real data.

of an one-hot encoded vector for the ground-truth of the i^{th} sample in a mini-batch, and $\hat{y}_{ij}$ corresponds to the j^{th} element of the output of the softmax FC layer for the i^{th} image.

Domain. We adopted the adversarial domain adaptation approach. In this work, domains are real and synthetic. A softmax FC layer for domain discrimination is added to the backbone network. The loss to make the network be trained indiscriminative to two domain is defined as follows:

$$L_{domain} = \frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i). \quad (3)$$

The domain discrimination loss is defined as the negative value of binary cross-entropy loss. Since the cross-entropy loss makes the network be trained discriminative between two domains, its negative loss would make the model more indistinguishable. If a vehicle captured by a camera is drawn by a graphic simulator in the same orientation, the features extracted from a synthetic image would be similar to that from a real image as the domain-dependent features are suppressed. The negative cross-entropy loss function is implemented by the gradient reversal layer [3].

Triplet Loss. In a mini-batch that contains P identities and Q images for each identity, each image (anchor) has $Q-1$ images of same identity (positives) and $(P-1)\times Q$ images of different identities (negatives). The triplet loss aims at pulling the positive pair (a, p) together while pushing the negative pair (a, n) away by a margin. That is, this loss forces the network to be trained to minimize the distance between the features from the same classes of images and, at the same time, to maximize the distance between the features from the different classes of images. The triplet loss[7] is defined as follows:

$$L_{tri} = \sum_{i=1}^{P}\sum_{a=1}^{Q}\left[m + \max_{p=1\ldots Q} D(v_{a,i}, v_{p,i}) - \min_{\substack{j=1\ldots P \\ n=1\ldots Q \\ j\neq i}} D(v_{a,i}, v_{n,j}) \right]_+ \quad (4)$$

where $v_{a,i}$ represents predicted vector of a^{th} image of the i^{th} identity group and m is the margin to control the difference between positive and negative pair distances and helps cluster the distribution more dense.

4.2. Disjoint Losses

Color, Type, and Orientation. The softmax cross-entropy loss is applied for these three targets. In fact, in terms of data type, orientation is continuous and of ratio type, whereas color and type are categorical and nominal. Therefore, it is natural to use regression to predict orientation. However, orientation estimation is one of the toughest problems for regression due to the wide range of the regression target. Actually, in our experiments, the optimization has not been converged for regression. Therefore, we convert the orientation regression to a direct classification into n discrete bins, with softmax cross-entropy loss, as shown in [35] or [5]. We divide the 360-degree orientation space into six bins of 60 degrees each. The cross-entropy losses for the color, type, and orientation are applied only to the synthetic images and set zero to the real images. The loss function can be presented as follows:

$$L_x = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\delta_i y_{ij}\log(\hat{y}_{ij}),\ \delta_i \in \{1, 0\}, \quad (5)$$

where x is one of color, type, and orientation, and δ_i is a mask value that is set to 1 if the i^{th} data in a mini-batch has x, and 0 otherwise.

5. Experiments

5.1. Evaluation Metric

To evaluate the performance of each model, we used the official evaluation metric for the AI City Challenge, which is the rank-K mean Average Precision (mAP) that measures the mean of average precision for each query considering only the top K matches. K is chosen to be 100. The average precision is computed for each query image by calculating the area under the Precision-Recall curve, and then the mean of the average precision over all the queries is computed.

5.2. Implementation

Our backbone network, ResNet-50, is initialized with the weights pre-trained on ImageNet [8] to accelerate the training process. We train the model end-to-end with an AMSGrad optimizer [20] for 60 epochs. The initial learning rate is set to 0.0003 and reduced by 0.1 after 20 and 40 epochs. The weight decay factor for L2 regulation is set to 0.0005, and the batch size is 64. For each mini-batch, two and two different vehicle-ids are selected from the real and synthetic datasets, respectively, and four images with the same ID are sampled. Therefore, a total of 16 different images with four different IDs from the real and synthetic datasets are sampled. An input image is resized to (128, 256). We adopt horizontal-flip and randomly-erase augmentations. In post-processing, we use the re-ranking algorithm proposed by Zhong et al. *et al.* [34], which is ordering the distance matrix between the features with the Jaccard distance and original output distance.

Case	O	C	T	V	D	Dataset	mAP
1				✓		R	25.48
2	✓			✓	✓	R+S	not converge
3		✓		✓	✓	R+S	35.16
4			✓	✓	✓	R+S	**38.35**
5	✓	✓		✓	✓	R+S	34.12
6	✓		✓	✓	✓	R+S	37.54
7		✓	✓	✓	✓	R+S	35.29
8	✓	✓	✓	✓	✓	R+S	33.96

Table 1. Results on the CityFlow-reID and VehicleX dataset for the 2020 AI City Challenge - Track 2. The results are from the official evaluation leaderboard. O, C, T, V, and D denote orientation, color, type, vehicle ID, and domain, respectively. Each box is checked if the target loss is included. In the dataset column, R and S represents real and synthetic data, respectively. The best result is bold.

Case	O	C	T	V	D	Dataset	mAP
1				✓		R	73.0
2	✓			✓	✓	R+S	**76.1**
3		✓		✓	✓	R+S	74.2
4			✓	✓	✓	R+S	74.9
5	✓	✓		✓	✓	R+S	74.7
6	✓		✓	✓	✓	R+S	75.3
7		✓	✓	✓	✓	R+S	74.8
8	✓	✓	✓	✓	✓	R+S	74.6

Table 2. Results on the VeRi and VehicleX dataset. O, C, T, V, and D denote orientation, color, type, vehicle ID, and domain, respectively. Each box is checked if the target loss is included. In the dataset column, R and S represents real and synthetic data, respectively. The best result is bold.

5.3. Results and Discussion

We trained and evaluated our models using the CityFlow-reID real dataset and the VeRI real dataset together with the synthetic data generated by VehicleX. The evaluation results of the models trained using the selected disjoint losses are shown in Table 1 and Table 2.

Performance on AI City Dataset. The baseline is Case 1 where a neural network is composed of the backbone network and the vehicle ID classifier. The baseline is trained with the real dataset using the vehicle-ID cross-entropy and triplet losses. As shown in Table 1, comparing with the baseline, the domain adaptation and semi-supervised learning approaches introduced in this study improve significantly the model performance by at least 8.48% (in Case

Method	mAP
FACT[12]	18.73
ABLN[38]	24.92
OIFE[29]	48.00
PROVID[16]	48.47
PathLSTM[22]	58.27
GSTE[]	59.47
VAMI[39]	61.32
BA[10]	66.91
BS[10]	67.55
PAMTRI[25]	71.88
StRDAN (baseline)	73.0
StRDAN (R+S, best)	**76.1**

Table 3. Comparing with other methods on VeRi dataset. Our StRDAN outperforms other methods.

8) and up to 12.87% (in Case 4). One interesting thing is that the model shows the best performance in Case 4 where only the vehicle type is considered among three labels of the synthetic image. On the contrary, in Case 8 where all three labels are considered, the model shows the worst performance.

Performance on VeRi Dataset.Table 2 also shows that the domain adaptation and semi-supervised learning approaches with synthetic dataset and additional losses contribute to performance improvement. The performance is improved by up to 3.1% in Case 2 and at least 1.2% in Case 3. Unlike the cases with the AI City dataset, the case with only the orientation label shows the best performance. However, in this case, the model cannot converge with the AI City dataset. In terms of performance, the models with Veri data is much better than those with AI City data. In table 3 we compare our StRDAN with other methods. Except for PAMTRI and StRDAN (R+S), all the models have been trained using only VeRi dataset. The table shows that our model outperforms the other methods in the table.

Domain Adaptation and Semi-supervised Learning. Based on the experimental results, it is clear that the domain adaptation and semi-supervised learning approaches contribute to extracting more important semantic features for vehicle Re-ID. However, there remains further research on unexpected phenomena: First, a model trained with only one loss out of three disjoint losses performs best. Second, the more disjoint losses are included, the lower the performance. Third, the best performance depends on the real dataset.

6. Conclusions

In this paper we propose an approach using domain adaptation and semi-supervised learning to fully utilize the synthetic data. Based on the experiment results, we found that increasing training data via with domain adaptation, improves performance. We also explored specific labels that only synthetic data has and discovered that using these labels with semi-supervised learning helps model extracting more semantic features.

As future work, the following issues need to be addressed.

- Synergy between disjoint losses and the real-world data dependency in the disjoint losses, which are discussed in previous section

- Effect of reality on synthetic data. The image data synthesized by VehicleX is easily distinguishable from real image data and very far from realistic. More realistic synthetic data obtained by driving simulation software can improve the performance much more.

- Prediction of orientation. We convert orientation regression to a direct classification into six discrete bins. However, as we have not tried various bin counts, it is necessary to investigate the optimal number of bins. Since the orientation is one of the key features to identify vehicles that are captured in various camera angles, the proper representation of orientation can boost performance.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *ArXiv*, abs/1412.4446, 2014.

[2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *ArXiv*, abs/1405.3531, 2014.

[3] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. *ArXiv*, abs/1409.7495, 2014.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *ArXiv*, abs/1505.07818, 2016.

[5] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ArXiv*, abs/1803.07728, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016.

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, abs/1703.07737, 2017.

[8] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

[9] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *CVPR Workshops*, 2019.

[10] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and P. Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2019.

[11] Chih-Ting Liu, Man-Yu Lee, Chih-Wei Wu, Bo-Ying Chen, Tsai-Shien Chen, Yao-Ting Hsu, Shao-Yi Chien, and NTU IoX Center. Supervised joint domain learning for vehicle re-identification. In *Proc. CVPR Workshops*, pages 45–52, 2019.

[12] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016.

[13] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016.

[14] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*, pages 869–884. Springer, 2016.

[15] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017.

[16] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20:645–658, 2018.

[17] Kai Lv, Heming Du, Yunzhong Hou, Weijian Deng, Hao Sheng, Jianbin Jiao, and Liang Zheng. Vehicle re-identification with location and time stamps. In *CVPR Workshops*, 2019.

[18] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, et al. The 2019 ai city challenge. In *CVPR Workshops*, 2019.

[19] Jinjia Peng, Huibing Wang, Tongtong Zhao, and Xianping Fu. Cross domain knowledge transfer for unsupervised vehicle re-identification. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 453–458. IEEE, 2019.

[20] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[21] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1900–1909, 2017.

[22] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1918–1927, 2017.

[23] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, et al. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 275–284, 2019.

[24] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, Shilei Wen, and Errui Ding. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *CVPR Workshops*, 2019.

[25] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William G. Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 211–220, 2019.

[26] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stanley T. Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8798, 2019.

[27] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[28] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 969–977, 2018.

[29] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 379–387, 2017.

[30] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. arXiv:1912.08855, 2019.

[31] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019.

[32] Yu Lin Zhang and Qiang Yang. A survey on multi-task learning. *ArXiv*, abs/1707.08114, 2017.

[33] Zhedong Zheng, Tao Ruan, Yunchao Wei, and Yezhou Yang. Vehiclenet: Learning robust feature representation for vehicle re-identification. In *CVPR Workshops*, 2019.

[34] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017.

[35] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2018.

[36] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6489–6498, 2018.

[37] Yi Zhou and Ling Shao. Vehicle re-identification by adversarial bi-directional lstm network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 653–662. IEEE, 2018.

[38] Yi Zhou and Ling Shao. Vehicle re-identification by adversarial bi-directional lstm network. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 653–662, 2018.

[39] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018.